**Analysis of Bronchoalveolar Lavage Fluid Metatranscriptomes Among Patients with COVID-19 Disease**

Michael Jochum (0000-0002-2398-356X) 1, Michael D. Lee (0000-0001-7750-9145) 2, Kristen Curry(0000-0001-8375-397X) 3, Viktorija Zaksas (0000-0001-7916-4902) 4,5, Elizabeth Vitalis (0000-0001-6036-1543)6, Todd Treangen (0000-0002-3760-564X) 3, Kjersti Aagaard (0000-0002-2960-0371) 1, Krista L. Ternus (0000-0003-1138-5308) 7

*1Department of Obstetrics and Gynecology, Division of Maternal-Fetal Medicine, Baylor College of Medicine and Texas Children’s Hospital, Houston, Texas 77030, USA; 2Blue Marble Space Institute of Science, Seattle, Washington 98104, USA; 3Department of Computer Science, Rice University, Houston, Texas 77005, USA; 4Center for Translational Data Science, University of Chicago, Chicago, Illinois 60615, USA; 5Clever Research Lab LLC, 2501 Chatham Rd, Suite N, Springfield, Illinois 62704, USA;* 6Inscripta*, Inc, 5500 Central Ave STE 220, Boulder, Colorado 80301, USA; 7Signature Science, LLC, 8329 North Mopac Expressway, Austin, Texas 78759, USA.*

**Author email addresses:**

Michael Jochum – [michael.jochum@bcm.edu](mailto:michael.jochum@bcm.edu)

Michael D. Lee – [mike.lee@nasa.gov](mailto:mike.lee@nasa.gov)

Kristen Curry – [kdc10@rice.edu](mailto:kdc10@rice.edu)

Viktorija Zaksas – [viktorija.zaksiene@gmail.com](mailto:viktorija.zaksiene@gmail.com)

Elizabeth Vitalis – [Elizabeth.Vitalis@inscripta.com](mailto:Elizabeth.Vitalis@inscripta.com)

Todd Treangen – [treangen@rice.edu](mailto:treangen@rice.edu)

Kjersti Aagaard – [aagaardt@bcm.edu](mailto:aagaardt@bcm.edu)

Corresponding Author: Krista L. Ternus – [kternus@signaturescience.com](mailto:kternus@signaturescience.com)

**Keywords:**

COVID-19, SARS-CoV-2, metatranscriptome, microbiome, bronchoalveolar lavage fluid, BALF, lower respiratory tract

**Abstract:**

Background

To better understand the potential relationship between COVID-19 and hologenome microbial community dynamics and functional profiles, we conducted a multivariate taxonomic and functional microbiome comparison of publicly available human bronchoalveolar lavage fluid (BALF) metatranscriptome samples amongst COVID-19 (*n*=32), community acquired pneumonia (*n*=25), and uninfected samples (*n*=29). We then performed a stratified analysis based on COVID-19 mortality amongst cohorts with known outcomes of deceased (*n=*10) versus survived (*n=*15). Our overarching hypothesis was that compared to uninfected and CAP metatranscriptome profiles, there are detectable and functionally significant relationships between bacterial metatranscriptomes associated with COVID-19 disease onset, progression, severity.

Results

Compared to the CAP and uninfected cohorts, we observed 34 (q < 0.05) functionally discriminant gene ontology (GO) terms in COVID-19 disease enriched in the COVID-19 disease cohort included hydrolase activity, and significant GO terms under the parental terms of biological regulation, viral process, and interspecies interaction between organisms. In our analysis stratifying based on COVID-19 mortality outcomes, we identified 21 GO terms including nucleobase-containing compound biosynthetic process, organonitrogen compound catabolic process, pyrimidine-containing compound biosynthetic process, and DNA recombination, RNA binding, magnesium and zinc ion binding, oxidoreductase activity, and endopeptidase activity. A Dirichlet multinomial mixtures clustering analysis resulted in a best model fit using three distinct clusters that were significantly associated with COVID-19 disease and mortality. We additionally observed discriminant taxonomic differences associated with COVID-19 disease and mortality in the genus *Sphingomonas,* belonging to the *Sphingomonadacae* family, *Variovorax,* belonging to the *Comamonadaceae* family, and in the class Bacteroidia*,* belonging to the order Bacteroidales.

Conclusions

To our knowledge, this study is the first to evaluate significant differences in taxonomic and functional signatures between BALF metatranscriptomes from COVID-19, CAP, and uninfected cohorts, as well as associating these taxa and microbial gene expression pathways with COVID-19 morbidity and mortality. Collectively, this study demonstrates significant relationships between the composition and functions of the human microbiome with COVID-19, CAP, and those shared or divergent between these two respiratory diseases states. The results from this study provide a strong scientific premise for testable hypotheses rendered herein, which warrant further investigation to provide a molecular understanding of the causality and directionality of host-microbiome-pathogen interactions.

**Background**

Metatranscriptomes from tissues and biologic samples arising from hosts with varying disease severity and outcomes represent a rich source of information to evaluate the role of the microbiome in onset and progression. For respiratory infections including SARS-CoV-2, bronchoalveolar lavage fluids (BALF) are collected to investigate the biology of lower respiratory tracts infections. Unfortunately, this sample type is more challenging to obtain for research studies that require large numbers of matching cases and controls, especially compared to the more easily accessible sample types like nasopharyngeal swabs. In general, BALF samples arise from patients that either have a clinical indication for them to be obtained or from healthy controls that have consented for the procedure. Early in the SARS-CoV-2 outbreak, scientists openly published metatranscriptome sequences from BALF of patients with COVID-19 disease; however, limitations in the sample numbers and lack of uniformity in study designs across different laboratories prevented a robust statistical analysis from taking place. In this paper, we computationally evaluate microbial insights drawn from these valuable BALF samples, despite the experimental study design limitations.In contrast to other studies that focus on characteristics of the human host response or SARS-CoV-2 lineages and viral variants, our analysis specifically evaluated the microbial taxonomic and functional profiles of the BALF metatranscriptomes. The role of the human microbiome in SARS-CoV-2 infection is poorly understood, but it remains important to study, since it could be a significant contributor to the observed variations in COVID-19 disease severity and resiliency between patients.

Among other risk factors, it is possible that the lower respiratory tract microbiome plays a role in COVID-19 disease severity. A previous 16S rRNA gene study found that COVID-19 patient endotracheal aspirates had lower microbial diversity compared to uninfected individuals, but these differences were not found to have a significant impact on fatality outcomes (1). The original Shen et al. study (2) performed a microbial taxonomic analysis of sequenced BALF metatranscriptomes without evaluating functional profiles or considering COVID-19 disease severity in the microbial analysis. Haiminen et al. (3) reanalyzed BALF metatranscriptome sequences from the Shen et al. study (2) and identified differences in expressed metabolic pathways in COVID-19 samples compared to the uninfected and CAP cohorts; however, functional profile differences were not analyzed based on COVID-19 clinical severity. Yang et al. (4) analyzed previously published BALF metatranscriptome datasets from multiple independent studies (2, 5-10) and performed a comparative taxonomic analysis between samples from COVID-19 patients and healthy control groups but did not subdivide cohorts further or perform functional analyses. Other studies have focused solely on the taxonomic analysis of a subset of published BALF metatranscriptomes and specific potential co-infections that may be present (11-12). To our knowledge, this study is the first to evaluate significant differences in taxonomic and functional signatures between BALF metatranscriptomes from COVID-19, CAP, and uninfected cohorts, as well as COVID-19 morbidity and mortality.

To better understand the potential relationship between COVID-19 morbidity and mortality and the human-microbiome, we conducted an analysis using human BALF metatranscriptome samples sourced from eight publications and nine corresponding public data repositories (Suppl. Tables 1-2) between the beginning of the COVID-19 pandemic to [March, 2021?]. BALF specimens from individual subjects grouped into one of three categorical classes: 1) uninfected controls; 2) community acquired pneumonia (CAP) patients; or 3) COVID-19 patients with moderate to severe disease, including death (Table 1). The objectives of the current study were to compare the BALF metatranscriptomes amongst and between each of the three cohort categorical classes or their sub-categories, such as COVID-19 severe disease versus death, and to identify significantly associated taxonomic and functional differences in microbial derived community dynamics. A primary objective was to rapidly assemble available metatranscriptomics data and generate a rigorous analysis pipeline for assessing host microbiome fitness, the composition of the microbiome taxa in association with respiratory disease, and the microbial gene expression pathways significantly perturbed. Our overarching testable hypothesis was that there is a potentially informative and discernably significant relationship between the BALF microbiome and the severity of COVID-19 disease.

Our overarching hypothesis was that there is a potentially informative and discernably significant relationship between the BALF microbiome and the severity of COVID-19 disease. We tested this hypothesis with the following objectives: (a) identify significantly associated taxonomic differences between each of the three cohort categorical classes (uninfected, CAP, or COVID-19), (b) discern microbiome-derived functional changes attributed to these community dynamics, and (c) assess these taxonomic and functional differences in relation to COVID-19 outcomes ranging from severe disease to death.

Table 1. Overview of Meta-analysis Dataset Clinical Characteristics (*n*=86).

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | Uninfected | Community Acquired Pneumonia | COVID-19 |
| **Cohort** | 29 (33.72%) | 25 (29.07%) | 32(37.21%) |
| **Outcome  (COVID-19 only)** |  |  |  |
| Deceased | - | - | 10 (31.25%) |
| Survived | - | - | 15 (46.87%) |
| Unspecified | - | - | 7 (21.88%) |
| **Sex** |  |  |  |
| female | 4 (18.18%) | 8 (36.36%) | 10 (45.45%) |
| male | 5 (13.15%) | 11 (28.94%) | 22 (57.89%) |
| unspecified | 20 (76.92%) | 6(23.07%) | 0 (0%) |
| **Reads** |  |  |  |
| paired | 29 (37.18%) | 25 (32.05%) | 24 (30.77%) |
| single | 0 (0%) | 0 (0%) | 8 (100%) |
| **Publication** |  |  |  |
| Chen | 0 (0%) | 0 (0%) | 2 (100%) |
| Ren | 9 (100%) | 0 (0%) | 0 (0%) |
| Shen | 20 (32.79%) | 25 (40.98%) | 16 (40.98%) |
| Wu | 0 (0%) | 0 (0%) | 1 (100%) |
| Xiong | 0 (0%) | 0 (0%) | 4 (100%) |
| Zhou | 0 (0%) | 0 (0%) | 9 (100%) |
| **Numeric variables (**mean ± SD) |  |  |  |
| Age | 53.2 ± 13.3 (*n=*9) | 51.2 ± 19.8  *(n=*17) | 47.3 ± 11.5  *(n=*32) |
| Temp. °C | - | 38.4 ± 0.91  *(n=*15) | 38.4 ± 0.715  *(n=*8) |
| days after onset | - | 9.07 ± 3.17  *(n=*14) | 12.05 ± 6.5  *(n=*41) |

**Methods**

*Data acquisition and exclusion.* Between the beginning of the COVID-19 pandemic [and March of 2021], we identified 8 studies with COVID-19 BALF samples. The publicly available Illumina reads were downloaded from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) or the China National Center for Bioinformation (CNCB) National Genomics Data Center (NGDC) (Suppl. Tables 1-2), along with the original publications where the clinical information was obtained for downstream analysis of BALF samples (2, 5-8, 13-15).

*Preprocessing*. Raw sequencing reads were assessed for quality (Phred>64) and adapters were trimmed with FastQC (16) and Trimmomatic (17), respectively. To control for different sequencing approaches by dataset (e.g., datasets being paired, or single-end reads), all paired-end reads were merged with FLASH (18) and concatenated with unmerged reads into one fastq file per sample. Human and PhiX reads were filtered out with a custom Kraken2 (19) built with solely human and PhiX references (see data and script availability section below), and low-complexity sequences were removed with fastp (20). Taxonomic analysis was subsequently performed with Kraken2 (19) utilizing their standard database.

*Microbiome taxonomic and functional analyses.* Parent-propagated GO term counts for all domains other than eukaryotes were imported into a working phyloseq (23) object alongside collected and curated clinical metadata using R 4.03 (24). GO term abundances from the remaining subjects’ specimens were compositionally transformed, center log ratio (CLR) normalized, and independently compared by case type (COVID-19 vs CAP and Uninfected) and survival outcome (COVID-19 only deceased vs survived) via MaAsLin2 (26) using minimum abundance, prevalence, and significance cutoffs of 0.01, 0.1, and q < 0.05 (Benjamini-Hochberg multiple test correction), respectively (27) (Suppl. Tables 3-4). Additionally, GO term counts subjected to unsupervised clustering community typing with Dirichlet Multinomial Mixtures (DMM) using square root scaled counts (28) (Suppl. Table 5), followed by statistical comparison using analysis of variance (ANOVA) with metadata categories case type and survival outcome. Statistically significant GO terms results derived from the MaAsLin2 analysis were thereafter ordered by parental lineage and visualized alongside consensus DMM clusters and metadata categories publication, case type, and survival outcome using the package pheatmap (v1.0.12) (29). Taxonomic differences were compared by case type and survival outcome with heat tree visualizations using log2 median ratio differences using metacoder (v0.34) (30).

*Data and code availability.* Overview of the processing workflow as well as all code used in the execution of the processing pipeline, analysis and visualization R scripts, and intermediate files have been made publicly available can be found online at the COV-IRT microbial GitHub repository (<https://github.com/COV-IRT/microbial>) and Open Science Framework (OSF) project (<https://osf.io/7nrd3/>) websites. The CoV-IRT-Micro conda package (<https://github.com/AstrobioMike/CoV-IRT-Micro>), along with programs modified from the *bit* package (22), was used to propagate parent GO terms, parse SeqScreen outputs by taxonomic domain, and summarize Kraken2 taxonomic results and SeqScreen-reported protein identifiers. SeqScreen matches reads to a reference database of proteins derived from UniProt; in cases where we track a specific GO term to its source proteins, we are referring to investigating what the underlying reference proteins to which our reads mapped. Additional information about the commands and versions of the tools used to process raw reads and assign taxonomies and GO terms can be found online on the Open Science Framework (OSF) project website (<https://osf.io/7nrd3/>).

**Results**

*Comparison between subject categorical classes (i.e., uninfected controls, or patients with Community Acquired Pneumonia (CAP) or COVID-19 disease)*

After controlling for random effects of publication and patient, results from the MaAsLin2 comparison across individual subjects were grouped by one of three categorical classes: 1) uninfected controls; 2) CAP patients; or 3) COVID-19 patients with moderate to severe disease, including death (Table 1). This revealed 35 out of 13,534 GO terms were associated with patients with COVID-19 when compared to patients with CAP or uninfected control subjects (Figure 1, Table 2, Table 3). Significant GO terms were grouped under seven parental GO terms (depth=1), including catalytic activity [GO:0003824], binding [GO.0005488], metabolic process [GO:0008152], cellular process [GO:0009987], biological regulation [GO:0065007], viral process [GO:0016032], and interspecies interaction between organisms [GO:0044419] (Figure 1, Suppl. Table 3).

Chart

Description automatically generated

Figure 1. Heatmap with Notable Microbially-Derived Gene Ontology Functional Annotations Associated with COVID-19 (n=32), as Compared to Community Acquired Pneumonia (n=29) & Uninfected (n=25) Cohorts. Rows are sorted by parental GO terms (depth=1), and columns are clustered by Euclidean distance using ward D2 clustering. Comparisons were conducted using MaAsLin2, controlling for publication and patient ID with Benjamini Hochberg multiple test comparison (q<0.05).

Table 2. MaAsLin2-Derived Significant Gene Ontologies Associated with COVID-19 (n=32) when Compared to the Community Acquired Pneumonia (n=25) Cohort. Comparisons were conducted using compositional transformed and CLR normalized count matrices, controlled for the random effects of publication and patient, and adjusted for multiple test comparisons using the Benajmini Hochberg correction method.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| feature | namespace | value (vs COVID-19) | coef | stderr | N | N.not.0 | pval | qval |
| hydrolase activity | GO:0016787 | Community Acquired Pneumonia | -0.008 | 0.001 | 86.000 | 69.000 | 0.000 | 0.000 |
| cellular process | GO:0009987 | Community Acquired Pneumonia | 0.010 | 0.002 | 86.000 | 74.000 | 0.000 | 0.000 |
| transferase activity | GO:0016740 | Community Acquired Pneumonia | 0.006 | 0.002 | 86.000 | 58.000 | 0.000 | 0.001 |
| modulation by virus of host cellular process | GO:0019054 | Community Acquired Pneumonia | -0.004 | 0.001 | 86.000 | 11.000 | 0.000 | 0.001 |
| modulation by symbiont of host cellular process | GO:0044068 | Community Acquired Pneumonia | -0.004 | 0.001 | 86.000 | 11.000 | 0.000 | 0.001 |
| biosynthetic process | GO:0009058 | Community Acquired Pneumonia | 0.005 | 0.001 | 86.000 | 23.000 | 0.000 | 0.001 |
| cellular macromolecule metabolic process | GO:0044260 | Community Acquired Pneumonia | 0.002 | 0.001 | 86.000 | 4.000 | 0.001 | 0.003 |
| organic substance biosynthetic process | GO1901576 | Community Acquired Pneumonia | 0.003 | 0.001 | 86.000 | 12.000 | 0.001 | 0.004 |
| cellular biosynthetic process | GO:0044249 | Community Acquired Pneumonia | 0.003 | 0.001 | 86.000 | 12.000 | 0.001 | 0.004 |
| transferase activity transferring phosphorus containing groups | GO:0016772 | Community Acquired Pneumonia | 0.005 | 0.001 | 86.000 | 14.000 | 0.001 | 0.004 |
| cellular metabolic process | GO:0044237 | Community Acquired Pneumonia | 0.004 | 0.001 | 86.000 | 72.000 | 0.005 | 0.012 |
| modulation by virus of host process | GO:0019048 | Community Acquired Pneumonia | -0.004 | 0.001 | 86.000 | 18.000 | 0.005 | 0.012 |
| nucleotidyltransferase activity | GO:0016779 | Community Acquired Pneumonia | 0.003 | 0.001 | 86.000 | 7.000 | 0.006 | 0.012 |
| metabolic process | GO:0008152 | Community Acquired Pneumonia | 0.005 | 0.002 | 86.000 | 76.000 | 0.006 | 0.013 |
| organonitrogen compound metabolic process | GO:1901564 | Community Acquired Pneumonia | 0.002 | 0.001 | 86.000 | 5.000 | 0.009 | 0.018 |
| modulation by symbiont of host process | GO:0044003 | Community Acquired Pneumonia | -0.003 | 0.001 | 86.000 | 18.000 | 0.018 | 0.032 |
| modulation of process of other organism involved in symbiotic interaction | GO:0051817 | Community Acquired Pneumonia | -0.003 | 0.001 | 86.000 | 18.000 | 0.018 | 0.032 |
| modulation of process of other organism | GO:0035821 | Community Acquired Pneumonia | -0.003 | 0.001 | 86.000 | 18.000 | 0.018 | 0.032 |
| organic substance metabolic process | GO:0071704 | Community Acquired Pneumonia | 0.003 | 0.001 | 86.000 | 76.000 | 0.026 | 0.045 |
| nucleic acid metabolic process | GO:0090304 | Community Acquired Pneumonia | -0.002 | 0.001 | 86.000 | 17.000 | 0.029 | 0.048 |

Table 3. MaAsLin2-Derived Significant Gene Ontologies Associated with COVID-19 (n=32) when Compared to the Uninfected (n=29) Cohort. Comparisons were conducted using compositional transformed and CLR normalized count matrices, controlled for the random effects of publication and patient, and adjusted for multiple test comparisons using the Benjamini Hochberg correction method.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| feature | namespace | value (vs COVID-19) | coef | stderr | N | N.not.0 | pval | qval |
| cellular process | GO:0009987 | Uninfected | 0.016 | 0.002 | 86.000 | 74.000 | 0.000 | 0.000 |
| metabolic process | GO:0008152 | Uninfected | 0.013 | 0.002 | 86.000 | 76.000 | 0.000 | 0.000 |
| modulation by symbiont of host cellular process | GO:0044068 | Uninfected | -0.007 | 0.001 | 86.000 | 11.000 | 0.000 | 0.000 |
| modulation by virus of host cellular process | GO:0019054 | Uninfected | -0.007 | 0.001 | 86.000 | 11.000 | 0.000 | 0.000 |
| modulation by virus of host process | GO:0019048 | Uninfected | -0.008 | 0.001 | 86.000 | 18.000 | 0.000 | 0.000 |
| organic substance metabolic process | GO:0071704 | Uninfected | 0.008 | 0.001 | 86.000 | 76.000 | 0.000 | 0.000 |
| cellular macromolecule metabolic process | GO:0044260 | Uninfected | 0.004 | 0.001 | 86.000 | 4.000 | 0.000 | 0.000 |
| cellular metabolic process | GO:0044237 | Uninfected | 0.009 | 0.001 | 86.000 | 72.000 | 0.000 | 0.000 |
| modulation by symbiont of host process | GO:0044003 | Uninfected | -0.008 | 0.001 | 86.000 | 18.000 | 0.000 | 0.000 |
| modulation of process of other organism | GO:0035821 | Uninfected | -0.008 | 0.001 | 86.000 | 18.000 | 0.000 | 0.000 |
| modulation of process of other organism involved in symbiotic interaction | GO:0051817 | Uninfected | -0.008 | 0.001 | 86.000 | 18.000 | 0.000 | 0.000 |
| hydrolase activity | GO:0016787 | Uninfected | -0.008 | 0.001 | 86.000 | 69.000 | 0.000 | 0.000 |
| interaction with host | GO:0051701 | Uninfected | -0.009 | 0.002 | 86.000 | 20.000 | 0.000 | 0.000 |
| viral process | GO:0016032 | Uninfected | -0.013 | 0.002 | 86.000 | 31.000 | 0.000 | 0.000 |
| transferase activity | GO:0016740 | Uninfected | 0.009 | 0.002 | 86.000 | 58.000 | 0.000 | 0.000 |
| primary metabolic process | GO:0044238 | Uninfected | 0.006 | 0.001 | 86.000 | 74.000 | 0.000 | 0.000 |
| symbiotic process | GO:0044403 | Uninfected | -0.014 | 0.002 | 86.000 | 31.000 | 0.000 | 0.000 |
| interspecies interaction between organisms | GO:0044419 | Uninfected | -0.014 | 0.002 | 86.000 | 31.000 | 0.000 | 0.000 |
| macromolecule metabolic process | GO:0043170 | Uninfected | 0.004 | 0.001 | 86.000 | 66.000 | 0.000 | 0.000 |
| organonitrogen compound metabolic process | GO:1901564 | Uninfected | 0.004 | 0.001 | 86.000 | 5.000 | 0.000 | 0.000 |
| binding | GO:0005488 | Uninfected | 0.004 | 0.001 | 86.000 | 81.000 | 0.000 | 0.001 |
| nitrogen compound metabolic process | GO:0006807 | Uninfected | 0.004 | 0.001 | 86.000 | 70.000 | 0.000 | 0.001 |
| biosynthetic process | GO:0009058 | Uninfected | 0.004 | 0.001 | 86.000 | 23.000 | 0.005 | 0.012 |
| ion binding | GO:0043167 | Uninfected | 0.002 | 0.001 | 86.000 | 8.000 | 0.006 | 0.012 |
| regulation of biological process | GO:0050789 | Uninfected | -0.003 | 0.001 | 86.000 | 15.000 | 0.010 | 0.020 |
| cellular nitrogen compound metabolic process | GO:0034641 | Uninfected | 0.002 | 0.001 | 86.000 | 53.000 | 0.011 | 0.021 |
| transferase activity transferring phosphorus containing groups | GO:0016772 | Uninfected | 0.004 | 0.001 | 86.000 | 14.000 | 0.014 | 0.027 |
| catalytic activity | GO:0003824 | Uninfected | 0.009 | 0.004 | 86.000 | 86.000 | 0.023 | 0.041 |
| RNA metabolic process | GO:0016070 | Uninfected | 0.002 | 0.001 | 86.000 | 6.000 | 0.028 | 0.048 |
| regulation of cellular process | GO:0050794 | Uninfected | -0.002 | 0.001 | 86.000 | 12.000 | 0.030 | 0.050 |

GO terms enriched in the COVID-19 cohort compared to the uninfected cohort included hydrolase activity [GO:0016787], as well as all significant GO terms with the parental terms of biological regulation [GO:0065007], viral process [GO:0016032], and interspecies interaction between organisms [GO:0044419]. Hydrolase activity [GO:0016787], nucleic acid metabolic process [GO:0090304], and many GO terms classified under interspecies interaction between organisms [GO:0044419] were also enriched in the COVID-19 cohort when compared to CAP. In contrast, GO terms enriched in the uninfected cohort compared to the COVID-19 cohort included all significant GO terms with the parental terms of cellular process [GO:0009987], metabolic process [GO:0008152], binding [GO.0005488], and terms classified under catalytic activity [GO:0003824] other than hydrolase activity [GO:0016787]. Results from the Dirichlet multinomial mixtures clustering analysis using all 13,534 gene ontology counts resulted in a best model fit using 3 distinct clusters that were significantly associated with each case cohort [p<0.0001] (Figure 1, Suppl. Table 5).

Taxonomic comparisons of the COVID-19 cohort to uninfected and CAP cohorts revealed a statistically significant decrease of several microbial genera within the phylum of Proteobacteria, including those of the families *Sphingomonadaceae* (i.e., *Sphingobium*, *Sphingopyxis*, *Sphingomonas*) and *Rhodobacteraceae* (i.e., *Paracoccus*) (Figure 2, Table 4). This included several species belonging to the genus *Sphingomonas* among BALF specimens from COVID-19 patients when compared to both the uninfected (p<0.0001, q <0.001) and CAP cohorts (p<0.005, q <0.05) (Suppl. Table 6). An analysis of the most common SeqScreen outputs taxonomically classified as *Sphingomonas* in BALF specimens among patients with COVID-19, irrespective of disease outcomes, included GO term assignments of hydrogen peroxide catabolic process [GO:0042744], response to oxidative stress [GO:0006979], catalase activity [GO:0004096], heme binding [GO:0020037], and metal ion binding [GO:0046872].

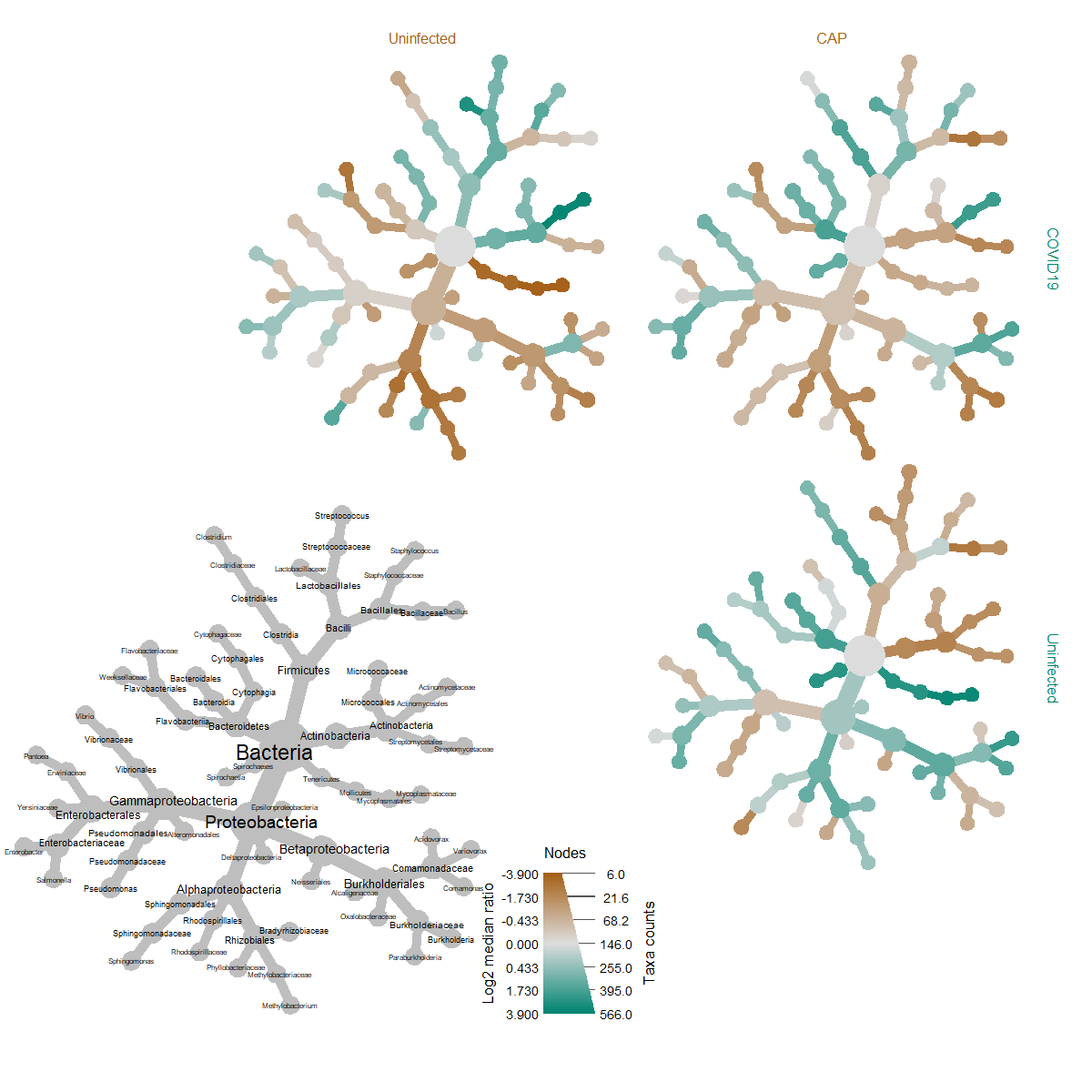


Figure 2. Heat Tree Matrix Visualizing Distinct COVID-19 vs. Uninfected & Viral Pneumonia Taxonomic Profiles. This taxonomic heat tree data matrix visualization depicts the log2 median ratio differences across the three different cohorts. The tree depicted in grey in the lower left represents a taxonomic assignment key for all the other trees. Each of smaller trees represents a comparison between the different cohorts, as labelled in the columns and rows. The taxa colored brown are more abundant among the cohort labelled in the columns, whereas taxa colored green are more abundant in the cohort labelled in the rows. For example, there were significant increases (green) identified in log2 median ratio of several species belonging to the genus *Sphingomonas* when comparing the COVID-19 to the uninfected cohort (top left) and decreases (brown) when comparing the uninfected cohort to the Community Acquired Pneumonia (CAP) cohort (bottom right).

Table 4. Log2 Median Ratio values of Top Taxa Associated with COVID-19 (n= 29) Compared to Community Acquired Pneumonia (n=25) and Uninfected (n=32) Cohorts. Values depicted in the table are mean values at the genus level for taxa that containing >10 species level significant comparisons with a qvalue < 0.05 and a log2 median ratio >1.0 using Wilcoxon rank sum test adjusted for multiple test comparison. A full list of significant taxonomic comparison values at the species level (when applicable) can be found in Supplementary Table 6.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| treatment 1 | treatment 2 | log2 median ratio | median diff | mean diff | p value | q value | taxon name |
| COVID-19 | Community acquired pneumonia | -2.0539 | -0.0728 | -0.1332 | 0.0067 | 0.0224 | *Paracoccus* |
| COVID-19 | Community acquired pneumonia | -2.0539 | -0.0728 | -0.1332 | 0.0067 | 0.0224 | *Sphingobium* |
| COVID-19 | Community acquired pneumonia | -2.0539 | -0.0728 | -0.1332 | 0.0067 | 0.0224 | *Sphingopyxis* |
| COVID-19 | Community acquired pneumonia | -2.0539 | -0.0728 | -0.1332 | 0.0067 | 0.0224 | *Sphingomonas* |
| COVID-19 | Uninfected | -4.3829 | -0.4587 | -0.3767 | <0.001 | <0.0001 | *Paracoccus* |
| COVID-19 | Uninfected | -4.3829 | -0.4587 | -0.3767 | <0.001 | <0.0001 | *Sphingobium* |
| COVID-19 | Uninfected | -4.3829 | -0.4587 | -0.3767 | <0.001 | <0.0001 | *Sphingopyxis* |
| COVID-19 | Uninfected | -4.3829 | -0.4587 | -0.3767 | <0.001 | <0.0001 | *Sphingomonas* |
| COVID-19 | Uninfected | -5.1294 | -0.4164 | -0.3356 | <0.001 | <0.0001 | *Bradyrhizobium* |
| COVID-19 | Uninfected | -5.1294 | -0.4164 | -0.3356 | <0.001 | <0.0001 | *Methylobacterium* |

*Metatranscriptomic comparison of BALF specimens from COVID-19 subjects sub-categorized and stratified by disease survival or death*

From subjects with known COVID-19 survival outcomes (i.e., of 32 samples, *n=*10 deceased, and *n=*15 survived) aperformed via MaAsLin2. After controlling for random effects of patient, we observed 21 unique GO terms which were significantly increased in their association with death or survival from COVID-19 disease (q-value < 0.05), with parental GO terms (depth=1) of metabolic process [GO:0090304], binding [GO.0005488], and catalytic activity [GO:0003824] (Table 5, Figure 3). GO terms with significant q-values (< 0.05) that were terminal in the observed GO term lineage (i.e., as specific as possible within the lineages of our result set), included nucleobase-containing compound biosynthetic process [GO:0034654], organonitrogen compound catabolic process [GO:1901565], pyrimidine-containing compound biosynthetic process [GO:0072528], and DNA recombination [GO:0006310] classified under the parental GO term of metabolic process [GO:0008152]; RNA binding [GO:0003723], magnesium ion binding [GO:0000287], and zinc ion binding [GO:0008270] classified under the parental GO term of binding [GO.0005488]; and oxidoreductase activity [GO:0016491] and endopeptidase activity [GO:0004175] classified under the parental GO term of catalytic activity [GO:0003824] (Suppl. Tables 4, 8-16).

Table 5. MaAsLin2-Derived Significant Gene Ontologies Associated with COVID-19 Disease Outcome (Deceased vs. Survived). Comparisons were conducted using compositional transformed and CLR normalized count matrices, controlled for the random effect of patient ID, and adjusted for multiple test comparisons using the Benajmini Hochberg correction method.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| name | ontology | namespace | depth | coef | stderr | pval | qval | N | N.not.zero |
| pyrimidine-containing compound metabolic process | biological\_process | GO:0072527 | 4 | -4.815 | 0.867 | <0.001 | <0.001 | 25 | 12 |
| nucleobase-containing compound biosynthetic process | biological\_process | GO:0034654 | 5 | -0.630 | 0.117 | <0.001 | <0.001 | 25 | 25 |
| transition metal ion binding | molecular\_function | GO:0046914 | 5 | -0.545 | 0.106 | <0.001 | <0.001 | 25 | 25 |
| aromatic compound biosynthetic process | biological\_process | GO:0019438 | 4 | -0.478 | 0.116 | <0.001 | 0.004 | 25 | 25 |
| heterocycle biosynthetic process | biological\_process | GO:0018130 | 4 | -0.393 | 0.100 | <0.001 | 0.007 | 25 | 25 |
| macromolecule biosynthetic process | biological\_process | GO:0009059 | 4 | 0.382 | 0.103 | <0.001 | 0.015 | 25 | 25 |
| RNA metabolic process | biological\_process | GO:0016070 | 6 | -0.310 | 0.086 | <0.001 | 0.018 | 25 | 25 |
| RNA phosphodiester bond hydrolysis | biological\_process | GO:0090501 | 7 | -1.412 | 0.402 | <0.001 | 0.024 | 25 | 17 |
| magnesium ion binding | molecular\_function | GO:0000287 | 5 | -2.336 | 0.709 | 0.001 | 0.036 | 25 | 11 |
| RNA binding | molecular\_function | GO:0003723 | 4 | 0.989 | 0.303 | 0.001 | 0.036 | 25 | 23 |
| zinc ion binding | molecular\_function | GO:0008270 | 6 | -0.880 | 0.266 | 0.001 | 0.036 | 25 | 24 |
| phosphorylation | biological\_process | GO:0016310 | 5 | 2.897 | 0.888 | 0.001 | 0.036 | 25 | 13 |
| organonitrogen compound catabolic process | biological\_process | GO:1901565 | 4 | -2.388 | 0.721 | 0.001 | 0.036 | 25 | 12 |
| endopeptidase activity | molecular\_function | GO:0004175 | 4 | -0.995 | 0.309 | 0.001 | 0.037 | 25 | 21 |
| pyrimidine-containing compound biosynthetic process | biological\_process | GO:0072528 | 5 | -5.505 | 1.711 | 0.001 | 0.037 | 25 | 7 |
| DNA recombination | biological\_process | GO:0006310 | 7 | -2.130 | 0.667 | 0.001 | 0.037 | 25 | 12 |
| oxidoreductase activity | molecular\_function | GO:0016491 | 2 | 2.541 | 0.801 | 0.002 | 0.037 | 25 | 13 |
| carbohydrate metabolic process | biological\_process | GO:0005975 | 3 | 2.245 | 0.717 | 0.002 | 0.039 | 25 | 15 |
| catalytic activity, acting on RNA | molecular\_function | GO:0140098 | 2 | -0.546 | 0.174 | 0.002 | 0.039 | 25 | 25 |
| pyrophosphatase activity | molecular\_function | GO:0016462 | 5 | -0.326 | 0.107 | 0.002 | 0.048 | 25 | 25 |
| organic cyclic compound binding | molecular\_function | GO:0097159 | 2 | 0.443 | 0.145 | 0.002 | 0.048 | 25 | 25 |
| hydrolase activity, acting on acid anhydrides | molecular\_function | GO:0016817 | 3 | -0.323 | 0.107 | 0.003 | 0.052 | 25 | 25 |

Chart

Description automatically generated

Figure 3. Heatmap of Significantly Different Gene Ontology Terms Associated with COVID-19 Mortality Comparing Deceased (n=10) versus Survived (n=15). Rows are sorted by parental GO terms (depth=1) and columns are clustered by Euclidean distance using ward D2 clustering. Comparisons were conducted using MaAsLin2, controlling for patient ID with Benjamini Hochberg multiple test comparison (q<0.05).

Of the nine terminal GO terms that were significantly different in this analysis (q-value < 0.05), RNA binding [GO:0003723] and oxidoreductase activity [GO:0016491] were the most enriched in samples from individuals that survived COVID-19 (Suppl. Table 4). An analysis of the proteins underlying the SeqScreen GO term assignments showed that RNA binding [GO:0003723] is being driven by an enrichment of 30S and 50S ribosomal proteins from the gram-positive cocci belonging to the genera *Streptococcus*, *Granulicatella*, *Enterococcus*, and *Lactococcus*, all of which were particularly prevalent in the “nCov7” survived COVID-19 patient from the Shen *et al.* study (Suppl. Table 7). The enrichment of the oxidoreductase activity [GO:0016491] term among survived COVID-19 patients was driven by many different samples and a variety of bacteria, including those from gram-positive bacteria belonging to the genera *Enterococcus*, *Streptococcus*, *Streptomyces*, *Pediococcus*, *Lactococcus*, and *Granulicatella*. Examples of underlying reference proteins to which reads mapped resulting in our observed oxidoreductase activity [GO:0016491] term included quinone oxidoreductase, pyruvate dehydrogenase, glyceraldehyde-3-phosphate dehydrogenase, and glyceraldehyde-3-phosphate dehydrogenase (Suppl. Table 13). Among the deceased COVID-19 patients, the terminal GO terms of endopeptidase activity [GO:0004175], zinc ion binding [GO:0008270], and nucleobase-containing compound biosynthetic process [GO:0034654] were being driven by an enrichment of SARS-CoV-2 proteins (Suppl. Tables 10, 12, 14). Mixed among proteins from other organisms, an enrichment of *Variovorax* proteins tagged with the terminal GO terms of pyrimidine-containing compound biosynthetic process [GO:0072528] (e.g., CTP synthase, putative sulfonate/nitrate transport system substrate-binding protein), organonitrogen compound catabolic process [GO:1901565] (e.g., histidine ammonia-lyase, aspartate/glutamate leucyltransferase), magnesium ion binding [GO:0000287] (e.g., phosphoribosyl-AMP cyclohydrolase, phosphopantetheinyl transferase), and DNA recombination [GO:0006310] (e.g., inclusive of possible *Variovorax* phage proteins - integrase family protein, putative transposase IS4 family, phage integrase family protein) appeared in the COVID-19 deceased patients. This enrichment of *Variovorax* proteins among samples from individuals who died of COVID-19 is consistent with the results from the taxonomic comparison analysis.

Compared to the survived group, the taxonomic comparisons in the deceased group revealed a statistically significant (p<0.0001, q <0.001) increase of the family *Comamonadaceae*, belonging to the genus *Variovorax*, and decreases in the family *Bacteriodales* (Figure 4, Table 6). There were no significant increases in taxa for CAP or COVID-19 individual analyses in Table 4, but in this analyses comparing COVID-19 outcome groups in Table 6, there majority of significant taxa changes were increases, while only 2 were decreased. [Go on to compare these results with the CAP/COVID-19 and COVID-19 specific taxa changes in figure 2 and table 4. Is there a taxa perturbed in COVID-19 death that was also associated with CAP? If not, you could say it is potentially COVID-19 death specific.]

Diagram

Description automatically generated

Figure 4. Heat Tree Demonstrating the BALF Metatransciptome Profiles Associated with COVID-19 Mortality. Notable increases were observed in the log2 median ratios in the Family *Comamonadaceae,* genus *Variovorax,* and significant decreases in the log2 median ratios of order *Bacteroidia* and class *Bacteroidales.*

Table 6. Log2 Median Ratio Counts of Taxa Associated with COVID-19 Mortality when Comparing Deceased (n=10) versus Survived (n=15). Comparisons were conducted using Wilcoxon rank sum test and adjusted for multiple test comparisons using the Benjamini Hochberg correction method.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| log2 median ratio | Median diff | Mean diff | p value | q value | Taxonomy |
| 2.25 | 0.361 | 0.371 | 0.00017 | 0.00691 | *Comamonadaceae* |
| 5.21 | 0.405 | 0.377 | 0.00017 | 0.00691 | *Variovorax* |
| 2.97 | 0.002 | 0.002 | 0.00353 | 0.074 | *Vibrionales* |
| 2.97 | 0.002 | 0.002 | 0.00353 | 0.074 | *Vibrionaceae* |
| 3.8 | 0.002 | 0.00181 | 0.00492 | 0.0827 | *Vibrio* |
| 1.84 | 0.0549 | 0.13 | 0.0137 | 0.124 | *Bacilli* |
| 2.24 | 0.403 | 0.297 | 0.0163 | 0.124 | *Burkholderiales* |
| 3.16 | 0.002 | 0.002 | 0.0157 | 0.124 | *Alteromonadales* |
| 3.61 | 0.004 | 0.004 | 0.0156 | 0.124 | *Yersiniaceae* |
| 2.1 | 0.005 | 0.00435 | 0.0156 | 0.124 | *Salmonella* |
| 1.77 | 0.011 | 0.064 | 0.0475 | 0.274 | *Streptococcaceae* |
| 2.29 | 0.425 | 0.296 | 0.0264 | 0.185 | *Betaproteobacteria* |
| -5.13 | -0.103 | -0.104 | 0.0308 | 0.199 | *Bacteroidia* |
| -5.18 | -0.099 | -0.102 | 0.00962 | 0.124 | *Bacteroidales* |

**Discussion**

Principal findings. To our knowledge, this study is the first to evaluate significant differences in taxonomic and functional signatures between BALF metatranscriptomes from COVID-19, CAP, and uninfected cohorts. Collectively, this study demonstrates significant relationships between the composition and functions of the human microbiome with COVID-19 and CAP, that are disease-specific, or shared. We then extend these analyses to identify the associations of these microbiome perturbations with COVID-19 mortality. Future studies will be aimed at assessing the changes in microbial species and genes responsible for causality or directionality of these host-microbiome interactions.

CAP-specific perturbations.

COVID-19-specific changes.

Divergence and similarities between CAP and COVID-19 metatranscriptomes. While our analyses focused on microbial taxa changes and microbe GO pathways, comparisons between the host and microbe gene expression changes may reveal synergistic or antagonistic responses. For instance, many of the CAP cases were associated with respiratory virus infections including rhinoviruses A, B, and C, influenza A, and B, and orthopneumovirus (Suppl. Table 1). The X terms and taxa that were X increased and X decreased, may reveal interactions between the host, bacterial species, and viruses that are intrinsic to respiratory virus responses, or divergent and specific to each pathogen. [taxa and GO specific lit review and discussions]

COVID-19 metatranscriptomics in the context of host immune responses. A subset of the studies from which the metatranscriptomics data we analyzed either primarily analyzed host gene expression[refs], or did not[refs]. [compare and contrast]

Microbiome changes associated with COVID-19 death. Distinct taxonomic features of BALF specimens from the COVID-19 vs. uninfected vs. CAP analysis included an increase in the genus *Sphingomonas,* belonging to the *Sphingomonadacae* family,among COVID-19 patients. Notable taxonomic features among COVID-19 patients with mortal disease included increases in log2 median ratios of the genus *Variovorax,* belonging to the *Comamonadaceae* family, and decreases in the class Bacteroidia*,* belonging to the order Bacteroidales. These findings are supported by previous reports associating *Sphingomonas* (31-34), a common opportunistic pathogen, with nosocomial infections. Among the COVID-19 cohort, one of the most highly expressed *Sphingomonas* genes was catalase [UniProt ID = J8VPL9]. This *Sphingomonas* catalase protein is assigned GO terms including hydrogen peroxide catabolic process [GO:0042744], response to oxidative stress [GO:0006979], catalase activity [GO:0004096], heme binding [GO:0020037], and metal ion binding [GO:0046872], and it is responsible for decomposing hydrogen peroxide into water and oxygen. This serves to protect cells from the toxic effects of hydrogen peroxide, which may suggest that *Sphingomonas* spp. respond to COVID-19 conditions in the patient by expressing genes that help them to survive well in environments undergoing great amounts of oxidative stress. Our findings additionally support a previous report regarding an increase in the abundance of *Variovorax* in COVID-19 patient BALF tissue (35). *Variovorax* spp. have also previously been reported in the microbiota of patients with lung cancer (36) and were shown to be a key driver of clustering amongst patients challenged with H1N1 influenza infections (37). The most abundantly expressed *Variovorax* proteins inthe COVID-19 cohort included those involved in cell wall organization and the plasma membrane (e.g., binding-protein-dependent transport systems inner membrane component [UniProt ID = E6VB76], endolytic peptidoglycan transglycosylase RlpA [UniProt ID = T1XG48]), oxidoreductase activity (e.g., methylenetetrahydrofolate reductase [UniProt IDs = J2L4W7, T1XH55], taurine dioxygenase [UniProt ID = T1XBI4], NADH-quinone oxidoreductase subunit H [UniProt ID = E6V509]), hydrolase activity (e.g., N-acyl-D-aspartate/D-glutamate deacylase [UniProt ID = J2T0U3], cytokinin riboside 5'-monophosphate phosphoribohydrolase [UniProt IDs = E6V0P4, J3CLH3]), and ATP-binding transport (e.g., ABC transporter related protein [UniProt ID = E6UUY9], extracellular solute-binding protein family 5 [UniProt ID = E6V3F7]).

Sexual dimorphisms. Based on our GO and taxa analyses, we found [or if not analyzed, include in limitations section below].

Strengths and limitations. While this work is rigorous and presents novel findings, this study is not without limitations. Data collection for this study ended in [March of 2020?]. Since then there have been additional COVID-19 BALF metatranscriptomics studies including bulk RNA-seq, scRNA-seq, and multiomics studies [refs]. A primary objective of this study was to generate a rapidly deployable analysis pipeline for emerging pathogen interactions with host microbiome fitness. At the time of this study [state the time frame again], the standard Kraken2 taxonomic database included the SARS-CoV-2 reference genome, but the SARS-CoV-2 proteins were not yet added to the SeqScreen database that was used for the functional analysis. This functional analysis demonstrated how GO terms and their corresponding proteins can be used to characterize an emerging pathogen (i.e., a pathogen that is not present in the reference database), as well as significant host microbiome functional shifts. SARS-CoV-2 reads were successfully detected in the taxonomic analysis of COVID-19 BALF samples, and GO terms associated with coronavirus proteins were found to be significantly different in the functional analysis. A number of coronavirus proteins were driving the significant associations of GO terms between COVID-19 and uninfected samples, including modulation by symbiont of host cellular process [GO:0044068], modulation by virus of host cellular process [GO:0019054], modulation by virus of host process [GO:0019048], modulation of process of other organism involved in symbiotic interaction [GO:0051817], modulation by symbiont of host process [GO:0044003], interaction with host [GO:0016032], viral process [GO:0051701], interspecies interaction between organisms [GO:0044419], modulation by symbiont of host cellular process [GO:0044068], and modulation by virus of host cellular process [GO:0019054] (Suppl Table 3). Coronavirus proteins were also driving notable GO term associations in COVID-19 deceased vs. survived, including transition metal ion binding [GO:0046914], zinc ion binding [GO:0008270], organic cyclic compound binding [GO:0097159], endopeptidase activity [GO:0004175], and nucleobase containing compound biosynthetic process [GO:0034654]. While samples from both COVID-19 deceased and survived individuals contained taxonomically and functionally classified coronavirus reads, the significant terminal GO terms of endopeptidase activity [GO:0004175], zinc ion binding [GO:0008270], and nucleobase-containing compound biosynthetic process [GO:0034654] were positively correlated with COVID-19 deceased patients. This was potentially due to a higher SARS-CoV-2 viral load, consistent COVID-19 mortality outcomes, and our finding that coronavirus proteins were tagged with these GO terms including the replicase polyprotein 1ab and 2'-O-methyltransferase (Figure X, Table X).

Of note, due to limitations in the depth of clinical metadata by subject, we could not distinguish between COVID-19 pathophysiology or associated medical comorbidities, treatments, nor interventions. However, because of the time interval in which COVID-19 patient specimens were recruited to their respective index studies at the beginning of the outbreak in Wuhan, China (i.e., 2019 and early 2020), COVID-19-specific interventions and treatments had yet to be introduced and thus comparisons between CAP and COVID-19 subject specimens would be less likely to be related to disease-focused therapy.

[comment on negative controls- kit negative, operator, and environmental controls]

[challenges of applying metagenomics analyses to metatranscriptomics data; propose a multiomics approach resolution]

**Conclusions**

COVID-19 disease has demonstrated a wide range of clinical severity outcomes, but the factors that correlate with disease severity are not fully understood. Here we rapidly assembled the available COVID-19 BALF metatranscriptomics data, created and utilized an analysis pipeline identifying significant taxonomic and functional differences associated with CAP caused by various pathogens, COVID-19, and compared shared and divergent metatranscriptomic perturbations associated with both of these diseases. By the nature of this analysis, this work does not address causality or directionality, especially for COVID-19 mortality. However, this work identifies taxonomic and functional relationships between the human microbiome and COVID-19, including putative taxa and functions associated with COVID-19 mortality. This study reveals numerous viral-bacterial-host interactions, rendering testable hypotheses with strong scientific premises warranting further investigation.

**Declarations**

*Ethics approval and consent to participate*: Not applicable

*Consent for publication*: Not applicable

*Availability of data and material*: The original sequence datasets used in this study were previously published and are publicly available in the locations described in Suppl. Tables 1-2. An overview of the data processing workflow, all code used in the execution of the processing pipeline, analysis and visualization R scripts, and intermediate files have been made publicly available can be found online at the COV-IRT microbial GitHub repository (<https://github.com/COV-IRT/microbial>) and Open Science Framework (OSF) project (<https://osf.io/7nrd3/> ; doi:10.17605/OSF.IO/7NRD3) websites. The OSF wiki (<https://osf.io/7nrd3/wiki/home/>) describes specific software tools and commands that were used to generate the results. The OSF project includes the following high-level project components relevant to this manuscript: Microbial\_Pre-Processing (i.e., outputs from quality trimming and filtering of raw sequence data), Metatranscriptome\_Kraken2 (i.e., Kraken2 taxonomic classification outputs), Metatranscriptome\_SeqScreen (i.e., SeqScreen final reports), Metatranscriptome\_GO\_Terms and Metatranscriptome\_GO\_Term\_Summaries (i.e., summaries of SeqScreen-assigned GO terms), and Metatranscriptome\_UniProt\_ID\_Counts (i.e., summaries of SeqScreen-assigned UniProt IDs). All methods were carried out in accordance with relevant guidelines and regulations. Suppl. Table 17 provides legends for all supplementary tables.

*Competing interests*: The authors declare that they have no competing interests.

*Funding*: The research performed by Drs. Ternus and Treangen for this study was partially funded by the Fun GCAT program from the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via the Army Research Office (ARO) under Federal Award No. W911NF-17-2-0089. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, ARO, or the US Government.

*Authors' contributions*: MJ statistically analyzed the taxonomic and functional results, created the figures, and wrote the manuscript. MDL developed code for sequence data analysis and provided reviews and guidance on how to analyze the data and generate figures. KC led the effort to run SeqScreen on computational resources provided at the Texas Advanced Computing Center. VZ reviewed data analysis results, performed background literature reviews, and contributed to writing the manuscript. EV assisted in analyzing the functional results and reviewing results in the context of prior studies. TT provided guidance in experimental design and computational analyses, as well as contributing to researching relevant prior publications. KA provided clinical expertise and reviewed the experimental design, data analysis results, and conclusions for clinical accuracy and relevance. KLT oversaw and contributed to the experimental design, data analysis, and writing of the manuscript. All authors read and approved the final manuscript.

*Acknowledgment*s: We would like to thank the COVID-19 International Research Team (COV-IRT) microbial subgroup team members and give special acknowledgment to John Fonner and the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing HPC resources that have contributed to the research results reported.

**References**

1. Merenstein C, Liang G, Whiteside SA, Cobián-Güemes AG, Merlino MS, Taylor LJ, Glascock A, et al. Signatures of COVID-19 severity and immune response in the respiratory tract microbiome. mBio. 2021;12(4): e0177721.

2. Shen Z, Xiao Y, Kang L, Ma W, Shi L, Zhang L, et al. Genomic diversity of severe acute respiratory syndrome–coronavirus 2 in patients with coronavirus disease 2019. Clinical Infectious Diseases. 2020;71(15): 713-720.

3. Haiminen N, Utro F, Seabolt E, Parida L. Functional profiling of COVID-19 respiratory tract microbiomes. Scientific Reports. 2021;11(1):6433.

4. Yang H, Zhilong J, Jinlong S, Weidong W, Kunlun H. The active lung microbiota landscape of COVID-19 patients through the metatranscriptome data analysis. BioImpacts. 2021; BI 12(2): 139–46.

5. Chen L, Liu W, Zhang Q, Xu K, Ye G, Wu W, et al. RNA based mNGS approach identifies a novel human coronavirus from two individual pneumonia cases in 2019 Wuhan outbreak. Emerg Microbes Infect. 2020;9: 313–319.

6. Zhou P, Yang X-L, Wang X-G, Hu B, Zhang L, Zhang W, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. 2020;579: 270–273.

7. Xiong Y, Liu Y, Cao L, Wang D, Guo M, Jiang A, et al. Transcriptomic characteristics of bronchoalveolar lavage fluid and peripheral blood mononuclear cells in COVID-19 patients. Emerg Microbes Infect. 2020;9: 761–770.

8. Michalovich D, Rodriguez-Perez N, Smolinska S, Pirozynski M, Mayhew D, Uddin S, et al. Obesity and disease severity magnify disturbed microbiome-immune interactions in asthma patients. Nat Commun. 2019;10: 5711.

9. Blanco-Melo D, Nilsson-Payant BE, Liu W, Uhl S, Hoagland D, Møller R, Jordan TX, et al. 2020. Imbalanced host response to SARS-CoV-2 drives development of COVID-19. Cell. 2020;181(5): 1036–45.e9.

10. Daamen AR, Bachali P, Owen KA, Kingsmore KM, Hubbard EL, Labonte AC, Robl R, Shrotri S, Grammar AC, and Lipsky PE. Comprehensive transcriptomic analysis of COVID-19 blood, lung, and airway. Scientific Reports. 2021;11(1): 7052.

11. Abouelkhair MA. Non-SARS-CoV-2 genome sequences identified in clinical samples from COVID-19 infected patients: evidence for co-infections. PeerJ 2020; 8(November): e10246.

12. Khan AA and Khan Z. COVID-2019-associated overexpressed Prevotella proteins mediated host-pathogen interactions and their role in coronavirus outbreak. Bioinformatics 2020;36(13): 4065–69.

13. Wu F, Zhao S, Yu B, Chen Y-M, Wang W, Song Z-G, et al. A new coronavirus associated with human respiratory disease in China. Nature. 2020;579: 265–269.

14. Huang W, Yin C, Wang G, Rosenblum J, Krishnan S, Dimitrova N, et al. Optimizing a metatranscriptomic next-generation sequencing protocol for bronchoalveolar lavage diagnostics. J Mol Diagn. 2019;21: 251–261.

15. Ren L, Zhang R, Rao J, Xiao Y, Zhang Z, Yang B, et al. Transcriptionally active lung microbiome and its association with bacterial biomass and host inflammatory status. mSystems. 2018;3.

16. Andrews S. FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (2015).

17. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014. pp. 2114–2120.

18. Magoč T, Salzberg SL. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics. 2011;27: 2957–2963.

19. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biol. 2019;20: 1–13.

20. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34: i884–i890.

21. Balaji A, Kille B, Kappell AD, Godbold GD, Diep M, Leo Elworth RA, et al. SeqScreen: Accurate and sensitive functional screening of pathogenic sequences via ensemble learning. bioRxiv. 2021. p. 2021.05.02.442344. doi:10.1101/2021.05.02.442344

22. Lee M. bit: a multipurpose collection of bioinformatics tools. F1000Research. 2022; 11:122. doi:10.12688/f1000research.79530.1

23. McMurdie PJ, Holmes S. phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data. PLoS One. 2013;8: e61217.

24. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2021. URL: https://www.R-project.org/

25. Davis NM, Proctor DM, Holmes SP, Relman DA, Callahan BJ. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. Microbiome. 2018;6: 1–14.

26. Mallick H, Rahnavard A, McIver LJ, Ma S, Zhang Y, Nguyen LH, et al. Multivariable association discovery in population-scale meta-omics studies. PLoS Computational Biology. 2021;17 (11): e1009442.

27. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful

approach to multiple testing. Journal of the Royal statistical society: series B (Methodological).

1995;57(1):289-300.

28. Holmes I, Harris K, Quince C. Dirichlet multinomial mixtures: generative models for microbial metagenomics. PLoS One. 2012;7: e30126.

29. Kolde R. pheatmap: Pretty heatmaps. 2018. https://cran.r-project.org/web/packages/pheatmap/

30. Foster ZSL, Sharpton TJ, Grünwald NJ. Metacoder: An R package for visualization and manipulation of community taxonomic diversity data. PLoS Comput Biol. 2017;13: e1005404.

31. Sirivongrangson P, Kulvichit W, Payungporn S, Pisitkun T, Chindamporn A, Peerapornratana S, et al. Endotoxemia and circulating bacteriome in severe COVID-19 patients. Intensive Care Med Exp. 2020;8: 72.

32. Chen S, Zhu Q, Xiao Y, Wu C, Jiang Z, Liu L, et al. Clinical and etiological analysis of co-infections and secondary infections in COVID-19 patients: An observational study. Clin Respir J. 2021;15: 815–825.

33. Ryan MP, Adley CC. Sphingomonas paucimobilis: a persistent Gram-negative nosocomial infectious organism. J Hosp Infect. 2010;75: 153–157.

34. Hsueh PR, Teng LJ, Yang PC, Chen YC, Pan HJ, Ho SW, et al. Nosocomial infections caused by Sphingomonas paucimobilis: clinical features and microbiological characteristics. Clin Infect Dis. 1998;26: 676–681.

35. Han Y, Jia Z, Shi J, Wang W, He K. The active lung microbiota landscape of COVID-19 patients through the metatranscriptome data analysis. BioImpacts. 2021;BI 12(2): 139–46.

36. Rose UD, Vetizou M, Roy S, Edwards B, Peck M, Smith L, et al. Role of the microbiota in primary lung cancer initiation and progression. The Journal of Immunology. 2019;202: 190.1–190.1.

37. Chaban B, Albert A, Links MG, Gardy J, Tang P, Hill JE. Characterization of the upper respiratory tract microbiomes of patients with pandemic H1N1 influenza. PLoS One. 2013;8.